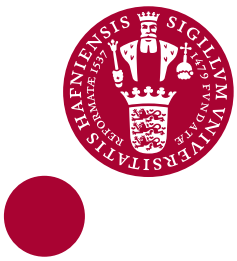


SignalP 5.0 improves signal peptide predictions using deep neural networks

Armenteros, José Juan Almagro; Tsirigos, Konstantinos; Sønderby, Casper Kaae; Petersen, Thomas Nordahl; Winther, Ole; Brunak, Søren; von Heijne, Gunnar; Nielsen, Henrik

Published in:
Nature Biotechnology

DOI:
10.1038/s41587-019-0036-z

Publication date:
2019

Document version
Peer reviewed version

SignalP 5.0 improves signal peptide predictions using deep neural networks

José Juan Almagro Armenteros[1,9], Konstantinos D. Tsirigos[1,2,3,4,9], Casper Kaae Sønderby[5], Thomas Nordahl Petersen[6], Ole Winther[5,7], Søren Brunak[1,8], Gunnar von Heijne[2,3] and Henrik Nielsen[1,*]

[1]Department of Bio and Health Informatics, Technical University of Denmark, Kgs Lyngby, Denmark
[2]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden.
[3]Science for Life Laboratory, Stockholm University, Solna, Sweden
[4]Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany
[5]Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[6]National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark
[7]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs Lyngby, Denmark
[8]Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[9]These authors have contributed equally to the presented work.

[*]Correspondence should be addressed to Henrik Nielsen (hnielsen@bioinformatics.dtu.dk)

Signal peptides (SPs) are short amino acid sequences in the N-terminus of many newly-synthesized proteins that target their passenger proteins for transfer across, or integration into, membranes. Previous tools have been used to predict SPs from amino acid sequence, but most cannot distinguish between specific types of signal peptides that occur in different organisms. Here, we present a deep neural network-based approach that improves SP prediction across all domains of life and distinguishes between three types of prokaryotic SPs.

Signal peptides (SPs) are found in a large number of nascent polypeptide chains in virtually all organisms and facilitate protein targeting to membrane-embedded export machineries in Bacteria[1], Archaea[2] and Eukaryotes[3]. SPs are found in secreted or transmembrane proteins, as well as proteins that reside inside specific organelles in eukaryotic cells. The general secretory pathway (Sec) is found in all domains of life and directs protein translocation across the plasma membrane in prokaryotes and the endoplasmic reticulum membrane in eukaryotes. Bacteria, Archaea, chloroplasts and some mitochondria also contain another major pathway termed Tat (Twin-Arginine Translocation). This pathway recognizes generally longer and less hydrophobic SPs that have a distinctive pattern usually containing two consecutive arginines (R-R) in the N-terminal region[4]. In contrast to the Sec pathway, which transports proteins in an unfolded state, the Tat pathway actively translocates folded proteins across the lipid membrane bilayer.

During or after translocation through the membrane, a signal peptidase removes the SP. For the most well-known type of SPs, this enzyme is called Signal Peptidase (SPase) I or

LepB in Bacteria, and orthologs of it are found in Archaea and in Eukaryotes, where it constitutes the only signal peptidase operating at the endoplasmic reticulum membrane. Bacterial lipoproteins are cleaved by a second signal peptidase, termed Signal Peptidase II or Lsp, which cleaves SPs that contain a conserved 'lipobox' in their C-terminus. This motif contains a cysteine immediately following the CS[5]. This cysteine site is found in both Bacteria and Archaea (although the actual SPase II has not been identified in Archaea[6]) and is vital to membrane anchoring[7]. Lastly, prokaryotic Type IV pilins are typically cleaved by a third signal peptidase, known as SPase III or prepilin peptidase (PilD) in bacteria and PibD in Archaea[8]. While bacteria contain Sec substrates that can be processed by SPase I, SPase II or SPase III, Tat substrates are only processed by SPase I or II.

Many algorithms to predict the presence of SPs and the location of their cleavage sites from amino acid sequences have been developed. SignalP was among the first publicly available methods[9], and has attracted a large number of users. Since then, several updates to this software have been published. While version 1[9] was based on artificial neural networks, version 2[10] additionally introduced hidden Markov models; version 3[11] enhanced the cleavage site predictions, and version 4[12] improved the discrimination between signal peptides and transmembrane helices. However, all previous versions have only been capable of predicting Sec-translocated SPs cleaved by SPase I; to predict Tat translocation or SPase II cleavage, more specialized tools were needed. Some of these methods are indeed appropriate only for one type of SPs (e.g. Sec/SPII or Tat/SPI) and thus cannot differentiate between all three classes (**Supplementary Table 1**).

Here, we present SignalP 5.0, a deep Neural Network-based method combined with Conditional Random Field classification and optimized transfer learning for improved SP prediction. The deep recurrent Neural Network architecture is better suited for recognizing sequence motifs of varying length, such as SPs, than traditional feed-forward Neural Networks. The Conditional Random Field imposes a defined grammar on the prediction and obviates the need for the post-processing step used in earlier versions of SignalP. Finally, the transfer learning makes it possible to obtain a decent performance also on small divisions of the dataset, notably the archaeal sequences (see **Online Methods** for details).

SignalP 5.0 distinguishes three types of signal peptides in prokaryotes: Sec substrates cleaved by SPase I (Sec/SPI), Sec substrates cleaved by SPase II (Sec/SPII), and Tat substrates cleaved by SPase I (Tat/SPI). SignalP 5.0 cannot identify Tat substrates cleaved by SPase II, although these are known to exist[13]. Nor have we been able to construct a sufficiently large dataset of SPase III-cleaved proteins for training a machine learning method and therefore our algorithm is not able to identify SPaseIII processed Sec substrates.

We trained and tested SignalP 5.0 on four groups of organisms (Eukaryotes, Archaea, Gram-positive bacteria, and Gram-negative bacteria) and four types of proteins: Sec/SPI, Sec/SPII, Tat/SPI and 'Other' (globular proteins without SP and transmembrane (TM) proteins with an experimentally verified TM segment within the first 70 amino acids

(AAs)). In total, the training data consisted of 20,758 proteins (**Supplementary Table 2**). After collecting the protein sequences, we clustered them using CD-HIT[14] at 20% sequence identity. The dataset was homology-partitioned into five sets. Each set had the same distribution of organisms and types of proteins. For each protein, the AAs were encoded using the normalized BLOSUM62[15] matrix such that each position in the sequence became a vector of length 20 containing the AA substitution probabilities. The labels for each AA were: Sec/SPI signal, Tat/SPI signal, Sec/SPII signal, outer region, inner region, transmembrane in-out, transmembrane out-in, SPI CS and SPII CS (**Online Methods**).

We benchmarked SignalP 5.0 against 18 SP prediction algorithms–as many as currently are available either as web-servers or standalone packages (**Supplementary Table 1**)**.** One particular method**,** Signal-BLAST[16], essentially does a BLAST[17] database look-up rather than a prediction from scratch and we observed that most of the proteins in our benchmark datasets were correctly predicted because they were identical to one of the proteins in Signal-BLAST's reference database. Because its performance therefore would be artificially high, Signal-BLAST was excluded from our benchmark (**Supplementary Note 1**)

Prediction performance for all SP detection algorithms was measured using the Matthews Correlation Coefficient (MCC)[18], where both true and false positive and negative predictions are counted at the sequence level. We used precision and recall to assess Cleavage Site (CS) predictions, where precision is defined as the fraction of CS

predictions that are correct, and recall is the fraction of real SPs that are predicted as the correct SP type and have the correct CS assigned.

SignalP 5.0 achieved an overall MCC of 0.938, 0.907, 0.890 and 0.966 for predicting Sec/SPI SPs for Archaea, Gram-negative bacteria, Gram-positive bacteria and Eukaryotes respectively. When tested on Sec/SPII SPs, SignalP 5.0 achieved MCCs of 0.956, 0.960 and 0.957 for Archaea, Gram-negative and Gram-positive bacteria, respectively. Finally, on Tat/SPI SPs, SignalP 5.0 had MCCs of 0.977, 0.981 and 0.868 for Archaea, Gram-negative and Gram-positive bacteria, respectively (**Supplementary Table 3**). In **Supplementary Table 4**, we demonstrate SignalP 5.0's discrimination performance on the different types of signal peptides with a confusion matrix, which shows the numbers of real and predicted examples in each class of sequences. Regarding CS precision, SignalP 5.0's performance varies between 0.630 and 0.970, whereas its CS recall varies between 0.579 and 0.970 (**Supplementary Table 5**).

To demonstrate the reliability of SignalP 5.0, we studied the probability distribution of correct and incorrect predictions. Prediction confidence was assessed by examining the probabilities of the most likely class predicted by the model from the AA sequences (**Supplementary Note 2 and Supplementary Fig. 1**).

A common problem in CS prediction is that experimental data used to train prediction algorithms can have erroneous or uncertain annotations. To account for this uncertainty, we considered a window of one, two and three AAs around the annotated CS position,

assuming that, if the annotation was incorrect, the correct position should be nearby. We reported a correct prediction if the predicted CS was within that window. The same was done for all other methods that were used in our benchmark (**Supplementary Note 2 and Supplementary Fig. 2**).

To construct an independent benchmark set for comparing the performance of SignalP 5.0 against all other prediction methods, we did a 20% homology reduction with CD-HIT between our training dataset and the dataset used for training the most recently published method, DeepSig[19] (which used SignalP4's training dataset). The result was a reduced benchmark dataset of 8,811 proteins (derived from the 20,758 proteins of the training dataset). While the benchmark set is independent with regards to eukaryotic and bacterial Sec/SPI data, this is not the case for the Sec/SPII, Tat/SPI and archaeal datasets, where many proteins were directly obtained from the training datasets of specialized predictors. In **Supplementary Table 2**, we report the constitution of the datasets for each organism type and category, both for training and for comparison.

Given that some methods were designed for a specific type of SPs and that not all methods run on all organism groups, we carried out three separate benchmarks (Sec/SPI, Sec/SPII and Tat/SPI SPs). Furthermore, because SignalP 5.0 is the only method capable of simultaneously predicting all types of SPs, each benchmark was run twice: first with only the respective SP type as 'positive' dataset and TM/Globular proteins as 'negative dataset' and then with adding the two remaining SP types to the 'negative' dataset. Importantly, the performance of SignalP 5.0 is measured on a cross-validated mode,

unlike the methods specialized for archaeal, Sec/SPII or Tat/SPI prediction, which contained some (or many) of the proteins of the benchmark already in their respective training datasets.

Benchmarks results are summarized in **Figs 1-2** and **Supplementary Tables 7-12**. SignalP 5.0 has the best SP discrimination across all organisms in the Sec/SPI benchmark, with the exception of Gram-positive bacteria where it ranks second after SignalP 4.1 (**Supplementary Table 7** and **Fig. 1**). It also has the highest CS recall in Eukaryotes and Bacteria, and the second highest CS recall in Archaea after PRED-SIGNAL[20], which, however, is a specialized method, trained on archaeal sequences only. Finally, Regarding CS precision SignalP 5.0 achieved the highest CS precision across all organisms compared with all existing methods (**Supplementary Table 8** and **Fig. 1**).

The performance of the otherwise quite successful methods Philius[21], Phobius[22] and SPOCTOPUS[23] on eukaryotic data was notably poorer than previously reported[12]. The reason for this discrepancy is that, in the current benchmark dataset, the number of eukaryotic SPs is much lower than in the previous study[12] (210 now versus 3,462 before), which makes the eukaryotic part of the evaluation dataset much more imbalanced than before. Finally, the performance of TOPCONS2[24], which is the only consensus method tested in our benchmark, is high in Bacteria, but not in Archaea or Eukaryotes, where it ranks below average. For CS predictions, it is clear that the consensus method is not ideal, but this was also not within the intended scope of this tool (**Supplementary Tables 7-8** and **Fig. 1**).

In the Sec/SPII SPs benchmark, SignalP 5.0 had superior performance across all metrics for all organisms, outperforming methods that were designed and optimized specifically for this this particular type of SPs (**Supplementary Tables 9-10** and **Fig. 2**).

SignalP 5.0 performs as well as PRED-TAT[25] and TATFIND[26] for predicting Tat/SPI SPs in Archaea and Gram-negative bacteria, although PRED-TAT has better prediction performance in Gram-positive bacteria. PRED-TAT achieved the highest CS recall in Bacteria, while SignalP 5.0 displayed the best CS prediction in Archaea. SignalP 5.0 demonstrated superior CS precision compared with PRED-TAT in Archaea and Gram-positive bacteria, although PRED-TAT achieved the highest CS precision in Gram-negative bacteria. TATFIND does not make CS predictions, and could therefore not be evaluated. When including Sec/SPI and Sec/SPII SPs in the 'negative' dataset, SignalP 5.0 performs as well as PRED-TAT in Archaea and has the highest CS prediction scores in Gram-negative and Gram-positive bacteria (**Supplementary Tables 11-12** and **Fig. 2**).

We used SignalP 5.0 to annotate two well-annotated reference proteomes; *Escherichia coli* (strain K12) and *Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (Baker's yeast). In both cases, SignalP 5.0 accurately detects all but one experimentally verified Sec/SPI SPs. The analysis also identified potentially new SPs, with high probability, which may be interesting candidates for verification (**Supplementary Note 3)**. Not only is SignalP 5.0 capable of predicting proteome-wide SPs across all organisms; it can also

classify them into Sec/SPI, Sec/SPII and Tat/SPI SPs, in most cases better than specialized predictors.

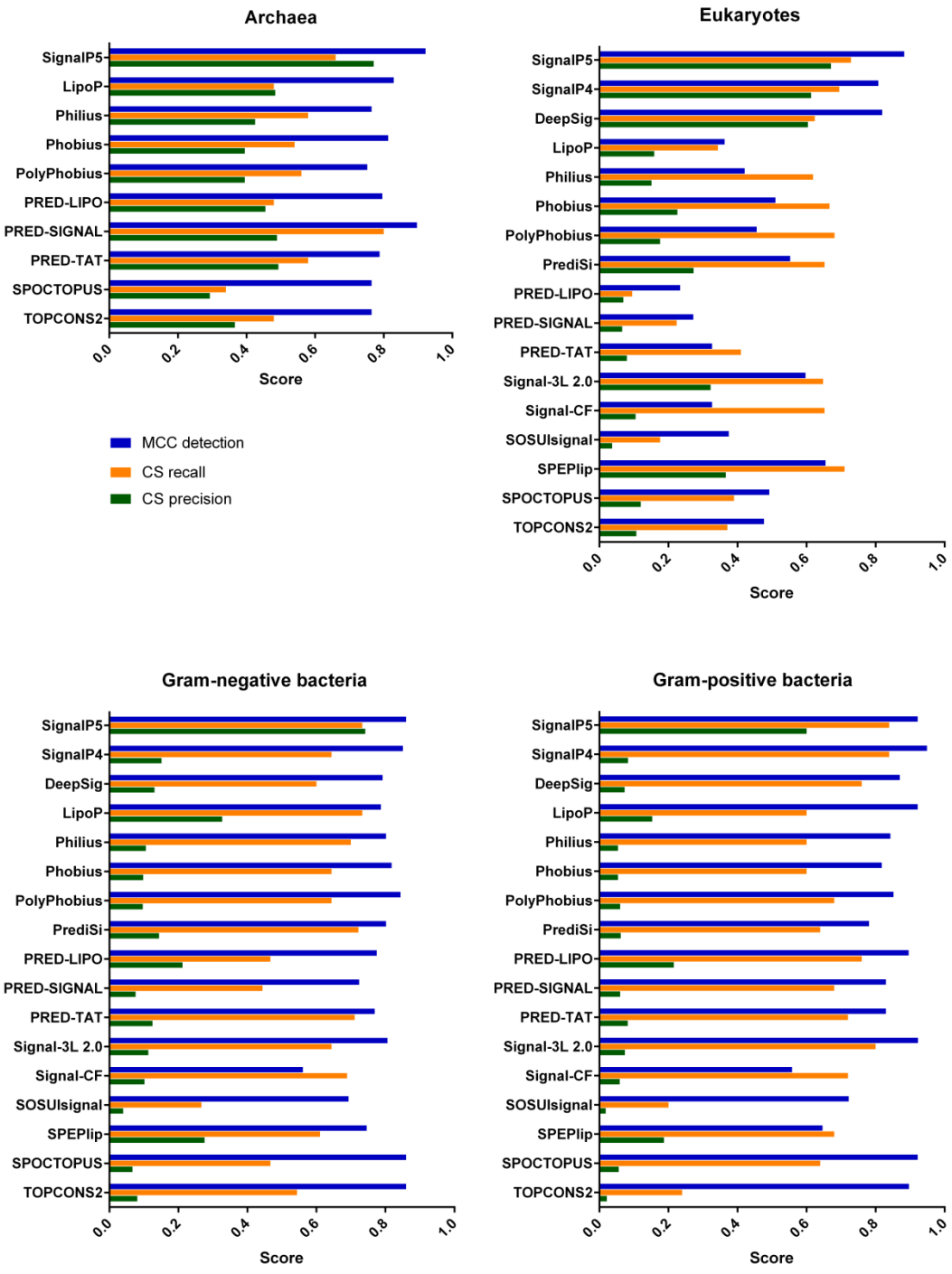


Figure 1. Performance comparison of signal peptide prediction algorithms across Archaea, Eukaryotes, Gram-negative bacteria and Gram-positive bacteria on Sec/SPI SPs. Performance was measured using the MCC detection and Cleavage Site recall and precision metrics.

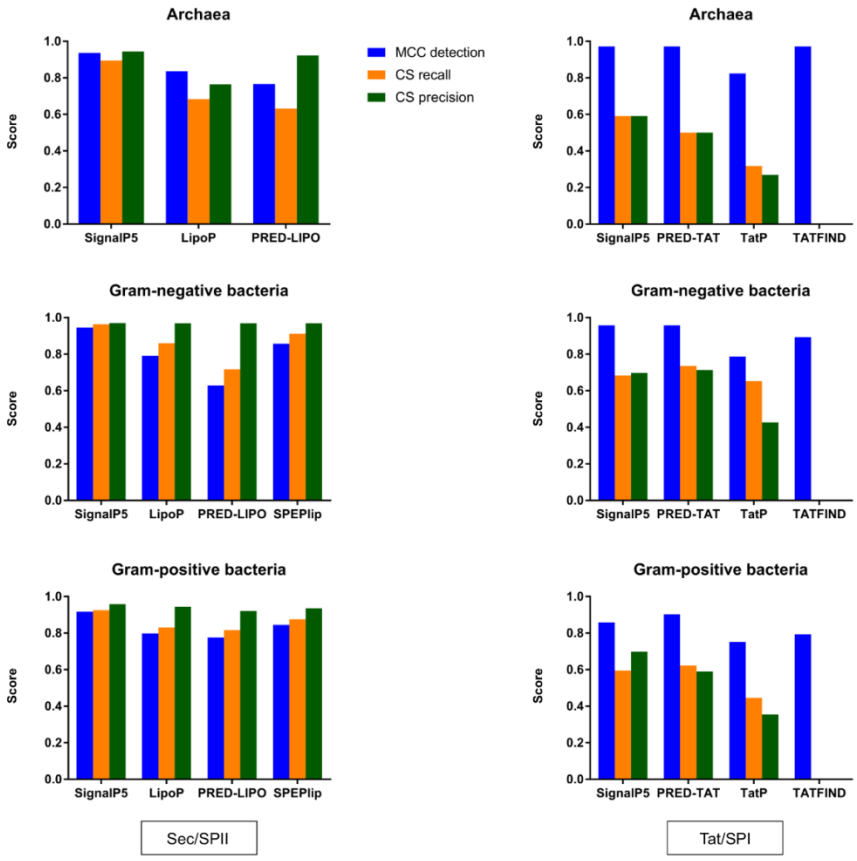


Figure 2. Performance comparison of signal peptide prediction algorithms on Sec/SPII and Tat/SPI substrates in Archaea, Gram-negative and Gram positive bacteria using the MCC detection and Cleavage Site recall and precision as metrics.

Acknowledgements

SB would like to acknowledge support from the Novo Nordisk Foundation (grant NNF14CC0001).

Author contributions

J.J.A.A designed the model architecture and trained the SignalP5 method with help from C.K.S. K.D.T collected the training and test datasets, performed the benchmarks and analyzed results. C.K.S, T.N.P, O.W., S.B and G.v.H provided suggestions during the design of SignalP5. K.D.T and H.N wrote the paper with input from J.J.A.A, C.K.S and O.W. H.N supervised and guided the project. All authors edited and approved the manuscript.

Competing financial interests
The downloadable version of SignalP 5.0 has been commercialized by the Technical University of Denmark (it is licensed for a fee to commercial users). The revenue from these commercial sales is divided between the program developers (J.J.A.A, K.D.T, C.K.S, T.N.P, O.W., S.B., G.v.H and H.N.) and the Technical University of Denmark.

Code availability
SignalP 5.0 is available at http://www.cbs.dtu.dk/services/SignalP/. The web version of SignalP 5.0 is free for all users, while the standalone package is free for academic users (and can be provided upon request), but is licensed for a fee to commercial users.

Data availability
The datasets used for training and testing SignalP 5.0 can be downloaded from http://www.cbs.dtu.dk/services/SignalP/data.php

References

1 Nouwen, N., Berrelkamp, G. & Driessen, A. J. *Journal of molecular biology* **372**, 422-433 (2007).
2 Pohlschroder, M., Gimenez, M. I. & Jarrell, K. F. *Current opinion in microbiology* **8**, 713-719 (2005).
3 Rapoport, T. A. *Nature* **450**, 663-669 (2007).
4 Berks, B. C. *Annual review of biochemistry* **84**, 843-864 (2015).
5 von Heijne, G. *Protein engineering* **2**, 531-534 (1989).
6 Pohlschroder, M., Pfeiffer, F., Schulze, S. & Halim, M. F. A. *FEMS microbiology reviews* **42**, 694-717 (2018).
7 Sankaran, K. & Wu, H. C. *The Journal of biological chemistry* **269**, 19701-19706 (1994).
8 Szabo, Z. et al. *Journal of bacteriology* **189**, 772-778 (2007).
9 Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. *Protein engineering* **10**, 1-6 (1997).
10 Nielsen, H. & Krogh, A. *Proceedings, International Conference on Intelligent Systems for Molecular Biology* **6**, 122-130 (1998).
11 Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. *Journal of molecular biology* **340**, 783-795 (2004).
12 Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. *Nature methods* **8**, 785-786 (2011).
13 Thompson, B. J. et al. *Molecular microbiology* **77**, 943-957 (2010).
14 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. *Bioinformatics* **28**, (2012).
15 Henikoff, S. & Henikoff, J. G. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-10919 (1992).
16 Frank, K. & Sippl, M. J. *Bioinformatics* **24**, 2172-2176 (2008).
17 Altschul, S. F. et al. *Nucleic acids research* **25**, 3389-3402 (1997).
18 Matthews, B. W. *Biochimica et biophysica acta* **405**, 442-451 (1975).

19 Savojardo, C., Martelli, P. L., Fariselli, P. & Casadio, R. *Bioinformatics* **34**, 1690-1696 (2017).
20 Bagos, P. G., Tsirigos, K. D., Plessas, S. K., Liakopoulos, T. D. & Hamodrakas, S. J. *Protein engineering, design & selection : PEDS* **22**, 27-35 (2009).
21 Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A. & Noble, W. S. *PLoS computational biology* **4**, e1000213 (2008).
22 Kall, L., Krogh, A. & Sonnhammer, E. L. *Journal of molecular biology* **338**, 1027-1036 (2004).
23 Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. *Bioinformatics* **24**, 2928-2929 (2008).
24 Tsirigos, K. D., Peters, C., Shu, N., Kall, L. & Elofsson, A. *Nucleic acids research* **43**, W401-407 (2015).
25 Bagos, P. G., Nikolaou, E. P., Liakopoulos, T. D. & Tsirigos, K. D. *Bioinformatics* **26**, 2811-2817 (2010).
26 Dilks, K., Rose, R. W., Hartmann, E. & Pohlschroder, M. *Journal of bacteriology* **185**, 1478-1483 (2003).

Online Methods

Sequence data

For eukaryotic and bacterial Sec/SPI signal peptides, we relied on the UniProt Knowledgebase release 2018_04[27]. Only reviewed entries (i.e. from UniProtKB/SwissProt) were used, and hypothetical proteins were not included. We discarded protein sequences shorter than 30 amino acids and we only considered signal peptides which had experimental evidence (ECO: 0000269) for the cleavage site. Gram-positive bacteria were defined as Firmicutes plus Actinobacteria. We did not include Tenericutes (*Mycoplasma* and related genera) since they do not seem to have a type I signal peptidase at all[28]. Gram-negative bacteria were defined as all other bacteria. All sequences were shortened to the 70 N-terminal AAs. For archaeal Sec/SPI signal peptides, we added the (few) experimentally verified proteins from SwissProt to the ones from the PRED-SIGNAL method.

For Tat/SPI signal peptides, we relied on a combination of the training set of the PRED-TAT method (which had experimentally verified Tat proteins for Gram-negative and Gram-positive bacteria) together with the 'Twin arginine translocation (Tat) signal profile' (PS51318) entry from the PROSITE database[29]. The status of PROSITE matches against Swiss-Prot entries is manually assessed during the curation process. Swiss-Prot curators evaluate the quality of the match according to the match score, the likelihood of the protein to contain such a domain and the appearance in other members of the protein family. The 'positive' status does not necessarily mean that the presence of the domain has been experimentally proven, but rather that the protein most probably contains such a domain according to the evaluation of the curators. For Archaea, we used proteins that were identified in the literature as being Tat/SPI (from the PRED-SIGNAL datasets) together with proteins belonging to the PS51318 entry of PROSITE.

A similar strategy was followed for the collection of the Sec/SPII dataset, where we used the respective PROSITE entry ('Prokaryotic membrane lipoprotein lipid attachment site profile' - PS51257), together with experimentally verified lipoproteins, taken from the PRED-LIPO[30] datasets. It should be noted that there was no overlap between the lists from these two PROSITE entries, i.e. we found no examples of proteins that belonged to both the Tat/SPII group.

We must stress here that, as it can be seen from **Supplementary Table 1**, the number of Sec/SPI signal peptides is relatively low as compared to the Tat/SPI and Sec/SPII types. This is due to the fact that, for Tat/SPI and Sec/SPII signal peptides, we relied mainly on

the PROSITE annotation and the PRED-TAT, PRED-SIGNAL and PRED-LIPO datasets, whereas, for Sec/SPI signal peptides, we used the annotation from UniProt. In 2014, UniProt adopted new evidence ontology. Before the change, an annotation was regarded as experimental if it lacked qualifiers such as 'Potential', 'Probable' or 'By similarity'; after the change, only annotations with a specific literature reference are annotated as being experimental (evidence code ECO:0000269). If we compare the number of experimental Sec/SPI SPs between the current version of SwissProt and the 2014_09 version (the last one before adopting the new scheme), then we observe 1,371 Eukaryotic, 280 Gram-negative and 118 Gram-positive Sec/SPI SPs that have lost their 'experimental' status (i.e. are missing the ECO: 0000269 annotation). SignalP 5.0 identifies 1,338/1,371 (97.59%) of the Eukaryotic Sec/SPI SPs and, of them 1,089 with a correct CS position (79.43%). In Gram-negative bacteria, the corresponding numbers are 222/280 (79.29%) for identification of Sec/SPI SPs and 195/280 (69.64%) for correct CS prediction. Finally, in Gram-positive bacteria, 105/118 (88.98%) of them were correctly identified, and 85/118 (72.03%) were found to have the same CS as the annotated one. These results are quite close to the overall performance of SignalP 5.0, indicating that these proteins could be correct SPs; however, we could not trust their experimental status, which is why we did not include them.

For transmembrane (TM) proteins, we relied on the TOPDB[31] database, which contains topological models of TM proteins based on either structural data (where there is interplay with the PDB_TM[32] database) or other experimental techniques, such as fusion with reporter enzymes, glycosylation studies, protease accessibility, immunolocalisation,

etc. If a TM protein was found to also contain an SP, then this protein was classified under the SP dataset.

Finally, we collected a globular proteins dataset, again from UniProt 2018_04, i.e. proteins with a subcellular location annotated as cytosolic (cytosolic, nuclear, mitochondrial, plastid, and/or peroxisomal in eukaryotes) and not belonging to the secretory pathway with experimental evidence (note that UniProt uses the term 'cytoplasm' for cytosol).

Methods for comparison

In addition to the previous version of SignalP (SignalP 4.1), 17 other methods were selected for comparison of predictive performances **Supplementary Table 2**. Most of the methods were downloaded and run locally on our computers or through their respective websites. For the methods Signal-3L 2.0[33], Signal-BLAST, Signal-CF[34] and SPEPlip[35], we wrote Perl scripts to automate the process of submitting a sequence and collecting the results. Signal-BLAST was eventually excluded from the benchmark (**Supplementary Note 1**).

SignalP 5.0 model architecture

SignalP 5.0 has three main novelties compared to previous versions: a) a powerful deep learning architecture[36] b) optimization using transfer learning[37] between multiple prediction tasks and c) Conditional Random Field (CRF) classification [38,39].

Deep Learning Model

The deep learning model (**Supplementary Fig. 3**), is composed of three primary components: 1) one-dimensional convolutions akin to learnable non-linear PSSMs, capturing short range correlations, 2) bidirectional Long-Short Term Memory (LSTM)[40] cells capturing long range sequence dependencies and 3) a Conditional Random Field (CRF) for predicting the class labels.

Transfer and multimodal learning

Deep learning models require relatively large amounts of data in order to train the models without overfitting and, as described above, we collected a dataset substantially larger than datasets previously used to successfully train Deep Learning models on protein sequences (see. e.g. Armenteros *et al*.[41] or Zhou & Troyanskaya[42]). However, some of the categories still had limited amounts of data available (**Supplementary Table 2**). To improve performance in organism groups with little data (notably Archaea), SignalP 5.0 utilizes transfer learning between taxonomic groups as well as multimodal learning predicting several related tasks using the same model. We trained a single unified model for Archaea, Gram-positive bacteria, Gram-negative bacteria and Eukaryotes, which improves performance on the low-data task since the model can learn generally useful features across all taxonomic groups. To inform the model about which taxonomic group a protein belongs to, we input an additional four-dimensional indicator vector into the LSTM cells of the model. To further improve performance, we employ multimodal learning, predicting both the individual amino acid labels as well as the global signal peptide type. Overall, transfer learning and multimodal learning means that, instead of

having eight models, each specialized to one group and one task, we use a single model that works for all the groups and performs both predictions at once.

Conditional Random Field

The CRF models a joint distribution of the sequential labels $y = y_1 \ldots y_T$ given the input sequence $x = x_1 \ldots x_T$ using the following restricted form:

$$p(y \mid x) = \frac{1}{Z(h)} = \prod_{t=1}^{T} \exp\left(\psi_{y_t}(h_t)\right) \prod_{t=1}^{T-1} \exp\left(\varphi_{y_t, y_{t+1}}\right)$$

where $h = h_1 \ldots h_T$ is the output of the core neural network model directly below the CRF, $Z(h)$ is the normalization constant of the distribution $p(y|x)$, $\psi(h_t)$ is a linear model which takes h_t as input and has the number of classes C outputs and φ_t is a trainable transition matrix with $C \times C$ parameters:

$$\psi(h_t) = W_\psi h_t + b_\psi$$

$$\varphi_{y_t, y_{t+1}} = W_\varphi$$

Due to the chain structure, inference can be carried out exactly using dynamic programming in $O(TC^2)$[43]. During training, where (x, y) is observed, we need to compute $Z(h)$ for each training sequence as part of the likelihood $p(y|x)$. During prediction, where only x is observed, we can calculate either the most probable sequence $argmax_y\ p(y|x)$ (using the Viterbi decoding algorithm) or the marginal probabilities $p(y_t|x)$, $t = 1, \ldots, T$. To make a single global prediction of whether a signal peptide is present or not in a protein, we take the average of the marginal probabilities across the sequence (nine classes: Sec/SPI signal, Tat/SPI signal, Sec/SPII signal, outer region, inner region, transmembrane in-out, transmembrane out-in, Sec SPI/Tat SPI cleavage site and Sec/SPII

cleavage site) and perform an affine linear transformation into four classes (Sec/SPI, Sec/SPII, Tat/SPI, Other), $l_s = W_s\left[\frac{1}{T}\sum_{t=1}^{T} p(y_t \mid x)\right]$, so as to get the logit of a categorical distribution over the presence or not of a signal peptide. To predict the location of the cleavage site, we use Viterbi decoding, since this ensures that a predicted sequence of signal peptide positions is always followed by prediction of a cleavage site.

Neural Network structure and optimization details

In this section, the neural network structure is described in more detail. The model is described sequentially going from the protein sequence input to predictions where the output of a layer is used as input for the next:

1. 1D convolution with 32 filters and a kernel width of three.
2. Bidirectional LSTM with 64 hidden units in the forward and backward models. To include the taxonomic group information in the model, a four-dimensional group indicator vector is concatenated to the input of the LSTM cells as illustrated in **Supplementary Fig. 3**.
3. 1D convolution with 64 filters and kernel widths five.
4. 1D convolution with nine filters (matching the number of position-specific classes) and kernel widths one.
5. Conditional Random Field for predictions. We calculate both the individual marginal probabilities of the labels at each position using the forward-backward algorithm and the global most likely label assignment for the entire sequence using Viterbi decoding. To predict the global label of the protein sequence, we

average the marginal probabilities across the sequence producing a 9x1 vector. We linearly map that vector to a 4x1 vector followed by a softmax function producing the global label prediction.

We used ReLu activation functions in all fully connected and convolutional layers and dropout was used to avoid overfitting[44]. The loss function consists of the sum of two terms, one for the individual amino acid label predictions and one for the global protein label prediction. Both terms are the cross-entropy between the predicted label distribution and the true observed label. All parameters were optimized using Stochastic Gradient Descent (SGD) on the loss function with mini batches of size 128 and a learning rate of 0.005. We optimized hyperparameters using Bayesian optimization[45] and five-fold nested cross-validation. The inner four folds were used to train four different models, each using three folds as training data and one fold as validation data, identifying thus the best set of hyperparameters (learning rate, LSTM hidden units, number of convolutional filters, convolutional filter width). After optimization, the fifth fold was used to assess the test performance and the procedure was repeated using each of the five folds as the test set. The advantage of this approach is that we obtain an unbiased test performance for the whole dataset at the expense of having to train 5x4 models. All the experiments were run using the Tensorflow library[46].

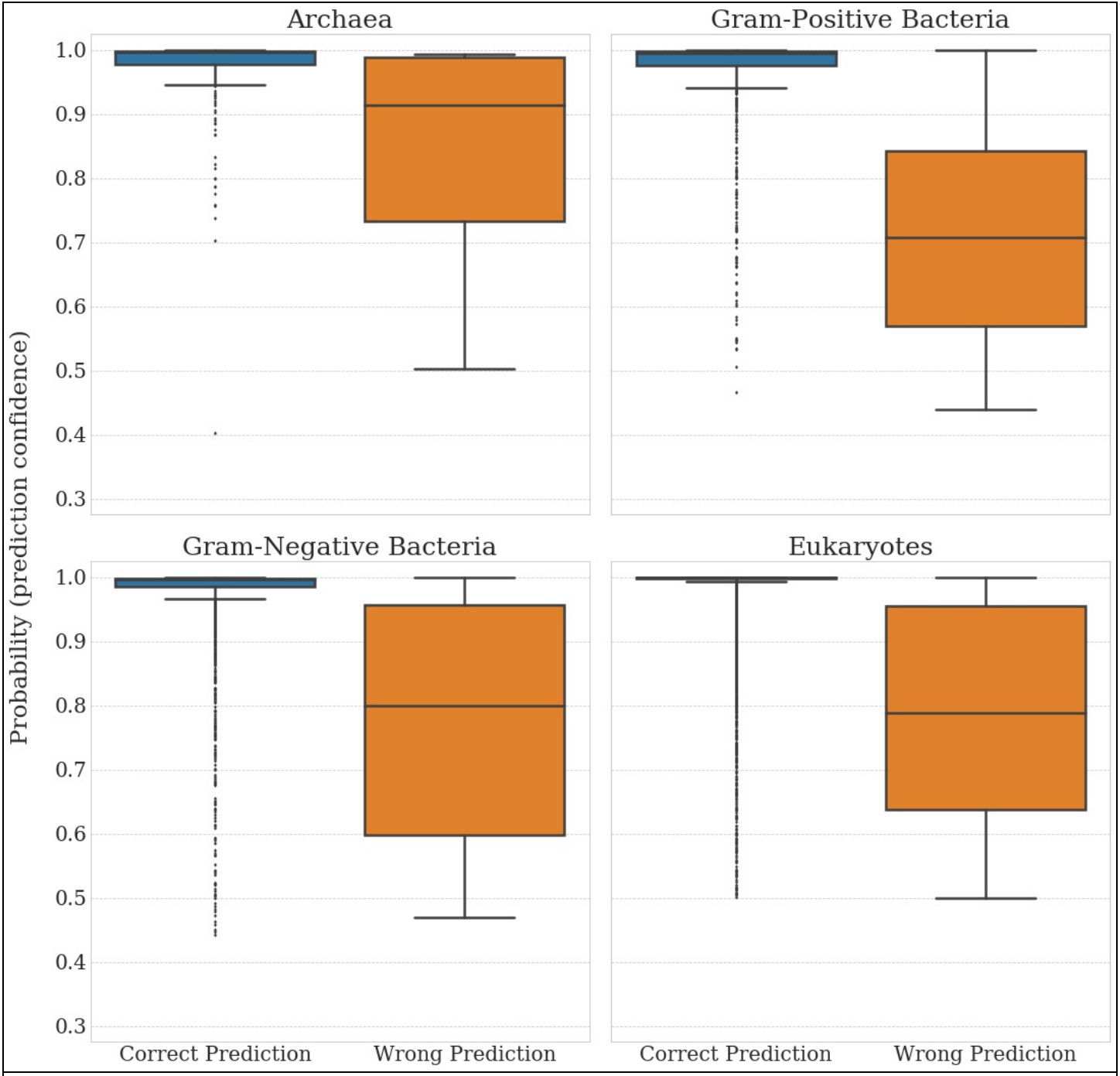
The incorporation of the aforementioned techniques improves the performance of SignalP 5.0 on both the signal type and the cleavage site prediction. **Supplementary Table 6** shows a comparison of the performance on the total training set (20,758 proteins) of the model using CRF and the organism group information, the model without the group

information and the model without CRF. The four models achieve similar performances, even though there are some differences worth mentioning. Regarding signal peptide detection, the CRF does not improve the performance considerably. However, its use is beneficial for the cleavage site prediction, where the difference in performance is significantly higher. Regarding the use of transfer learning, it is clear that the models without this feature are the ones with the worst performance on the signal peptide detection. For instance, on archaeal proteins, the model has a performance of 0.913, while, for the model that utilizes transfer learning, the performance climbs to 0.966.

Online references

27 UniProt Consortium, *Nucleic acids research* **46**, 2699 (2018).
28 Fraser, C. M. et al. *Science* **270**, 397-403 (1995).
29 Sigrist, C. J. et al. *Nucleic acids research* **41**, D344-347 (2013).
30 Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D. & Hamodrakas, S. J. *Journal of proteome research* **7**, 5082-5093 (2008).
31 Dobson, L., Lango, T., Remenyi, I. & Tusnady, G. E. *Nucleic acids research* **43**, D283-289 (2015).
32 Kozma, D., Simon, I. & Tusnady, G. E. *Nucleic acids research* **41**, D524-529 (2013).
33 Zhang, Y. Z. & Shen, H. B. *Journal of chemical information and modeling* **57**, 988-999 (2017).
34 Chou, K. C. & Shen, H. B. *Biochemical and biophysical research communications* **357**, 633-640 (2007).
35 Fariselli, P., Finocchiaro, G. & Casadio, R. *Bioinformatics* **19**, 2498-2499 (2003).
36 LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436-444 (2015).
37 Pan, S. J. & Yang, Q. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345 - 1359 (2010).
38 Lafferty, J. D., McCallum, A. & Pereira, F. C. N. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 282-289 (2001).
39 Hochreiter, S. & Schmidhuber, J. *Neural Computation* **9**, 1735-1780 (1997).
40 Graves, A. *Supervised Sequence Labelling with Recurrent Neural Networks* (2012).
41 Almagro Armenteros, J. J., Sonderby, C. K., Sonderby, S. K., Nielsen, H. & Winther, O. *Bioinformatics* **33**, 3387-3395 (2017).
42 Zhou, J. & Troyanskaya, O. G. *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 745-753 (2014).

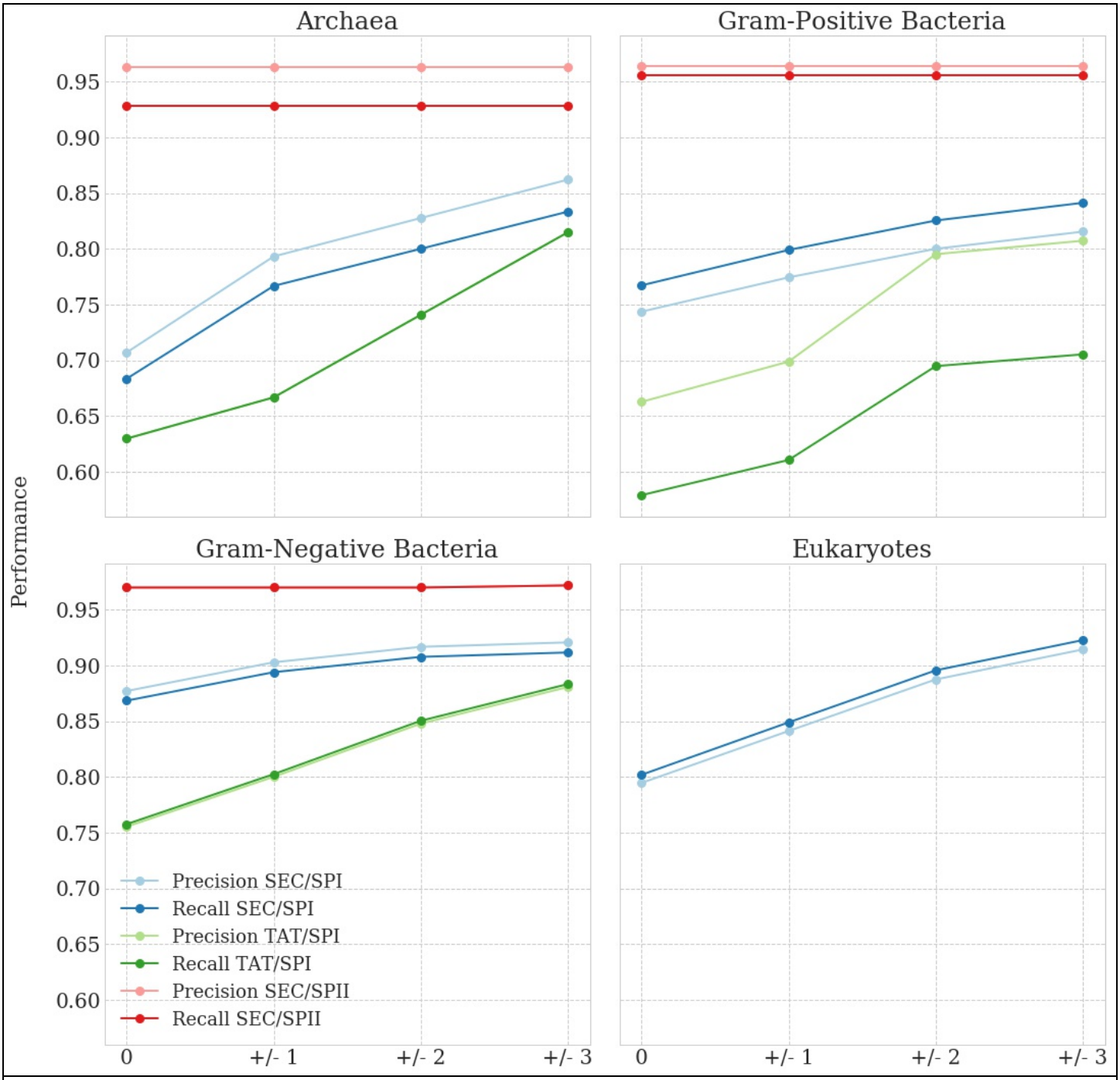
43 Bishop, C. *Pattern recognition and machine learning* (2007).
44 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. *The Journal of Machine Learning Research* **15**, 1929-1958 (2014).
45 Hutter, F., Hoos, H. H. & Leyton-Brown, K. *Proceedings of the 5th international conference on Learning and Intelligent Optimization*, 507-523 (2011).
46 Abadi et al. *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, 265-283 (2016).
47 Juncker, A. S. et al *Protein science* **12**, 1652-1662 (2003).
48 Kall, L., Krogh, A. & Sonnhammer, E. L. *Bioinformatics* **21**, i251-257 (2005).
49 Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. *Nucleic acids research* **32**, W375-379 (2004).
50 Gomi, M., Sonoyama, M. & Mitaku, S. *Chem Bio Informat J* **4**, 142-147 (2004).
51 Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T. & Brunak, S. *BMC bioinformatics* **6**, 167-173 (2005).



Supplementary Figure 1

Boxplot of the probability of the predicted class for correct and incorrect predictions

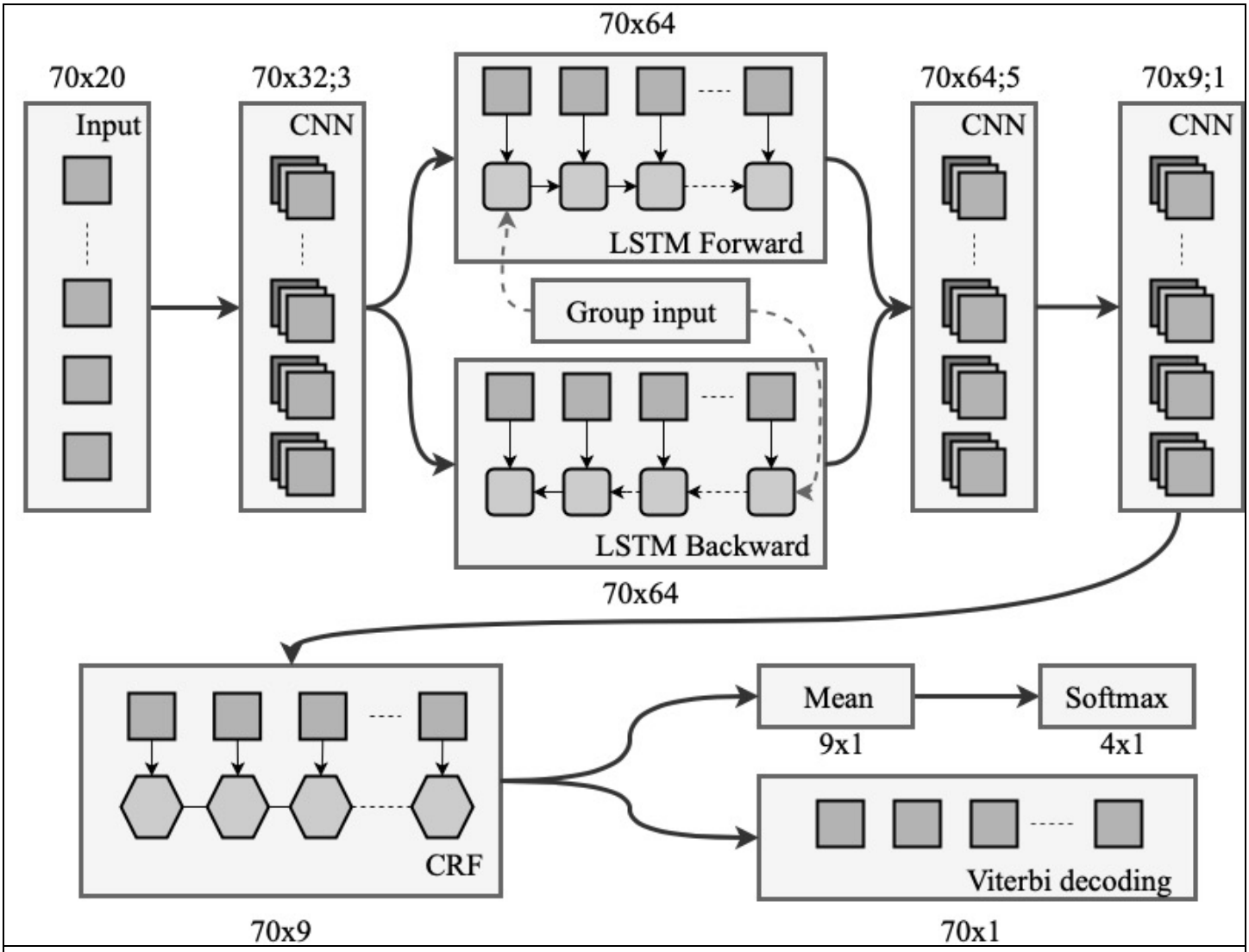
A probability close to 1 means a highly reliable prediction. For Archaea, Gram-Positive and Gram-Negative the probability threshold is 0.25, as there are four possible classes (*Sec/SPI*, *Tat/SPI*, *Sec/SPII* and *Other*). For Eukarya this threshold is 0.5, as it has only two classes (*Sec/SPI* and *Other*). A probability close to this threshold means a very unreliable prediction. All classes, namely *Sec/SPI*, *Tat/SPI*, *Sec/SPII* and *Other* are combined in this plot.



Supplementary Figure 2

Performance of SignalP 5.0 on cleavage site detection when considering a window of 0, 1, 2 and 3 amino acids centered on the real cleavage site.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary. To ensure accurate appearance in the published version, please use the Symbol font for all symbols and Greek letters.



Supplementary Figure 3

The SignalP 5.0 Neural Network architecture.

Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary. To ensure accurate appearance in the published version, please use the Symbol font for all symbols and Greek letters.

Supplementary Material

Supplementary Note 1

The special case of Signal-BLAST

Signal-BLAST runs BLAST against a pre-constructed database of signal peptides, and, if it finds a hit with high similarity, it essentially conducts a database look-up (current version of Signal-BLAST uses Uniprot 2017_6 as reference). Because of that, we sought to find out how many of the proteins in our datasets fall under the "look-up" category. Interestingly, out of the 7,457 proteins in Eukaryotes, modes (parameter sets) 1, 2 and 3 of Signal-BLAST find 7,444 /7,457 (99.8%) with 100% identical sequence in the reference database, and only mode 4 is different with 3,160 /7,457 (42.4%). In this benchmark, the overall MCC of modes 1, 2 and 3 was almost equal to 1.0, while for the mode 4 it merely reached 0.145. In Archaea, out of 182 proteins, there are 126 (for modes 1, 2 and 3) or 109 (from mode 4) proteins with a 100% identity, which translates into 69.2% or 59.8% of the total proteins in this set. In Gram-negative bacteria, out of 783 proteins, there are 723 (for modes 1, 2 and 3) or 684 (from mode 4) proteins with a 100% identity in the reference database (constituting 92.3% or 87.4% of them). Finally, in Gram-positive bacteria, out of 389 proteins, there are 343 (for modes 1, 2 and 3) or 324 (from mode 4) proteins with 100% identity to one of the proteins in the reference database, i.e. 88.2% or 83.2% of them. Based on this finding, the Signal-BLAST method was not included in our benchmark, since its predictive performance would be artificially high. Signal-BLAST further exemplifies that the idea of using sequence similarity alone is probably not suited for the task of SP prediction, since, if no homology is found, then the prediction performance is very low.

Supplementary Note 2

Signal peptide classification and Cleavage Site prediction reliability

We observed that correctly classified proteins have a prediction confidence close to 1, even though some outliers were present. The median of these probabilities for Archaea, Gram-positive bacteria, Gram-negative bacteria and Eukaryotes is 0.996, 0.995, 0.997 and 0.999 respectively. In contrast, incorrectly classified proteins have wider probability distributions, ranging between 0.4 and 1.0. The median of these probabilities for Archaea, Gram-Positive, Gram-Negative and Eukaryotes is 0.914, 0.707, 0.800 and 0.788, respectively, highlighting the lower prediction confidence of incorrect predictions.

CS prediction performance for Tat/SPI and Sec/SPI proteins increased across all domains in relation to window size (**Supplementary Figure 2**). For Eukaryotes, recall increases from 0.802 to 0.923 when considering a window of three amino acids, indicating that a high proportion of the erroneous predictions are within 3 AAs of the annotated CS position. We can therefore conclude that, when predicting the SPase I CS with SignalP 5.0, we can be highly confident that the real CS will be located in a window of three AAs around the predicted CS. For Sec/SPII SPs, the cleavage site performance is very high and is not improved by considering a wider window, reflecting that the SPase II CS is dependent on the presence of a cysteine residue.

Supplementary Note 3

Proteomic analysis using SignalP 5.0

For *Escherichia coli,* UniProt reports 498 proteins with a SP, 137 of which are experimentally verified and 361 predicted by various ways. Out of the 137 SPs, 122 are Sec/SPI, 11 are Sec/SPII and 4 are Tat/SPI SPs. SignalP 5.0 predicts 612 SPs in total, out of which 414 Sec/SPI, 161 Sec/SPII and 37 Tat/SPI. When we compare the UniProt annotated SPs to the SignalP 5.0 predicted ones, we observe that SignalP 5.0 detects 136/137 experimental SPs, together with their respective type (misses one Tat/SPI SP). Regarding CS prediction, 131/137 are predicted to have precisely the same CS, whereas, 5/137 are within the ±3 window.

The respective analysis for *Saccharomyces cerevisiae* is as follows: UniProt annotation reports 297 proteins, 37 of which are experimentally annotated and 260 are predicted by various ways. SignalP 5.0 predicts 314 Sec/SPI SPs in total. When we compare the UniProt annotated SPs to the SignalP 5.0 predicted ones, we observe that SignalP 5.0 detects 36/37 experimental SPs. Regarding CS prediction, 29/37 are predicted to have precisely the same CS, while, a further of 2/37 are within the ±3 window.

These two analyses show us that SignalP 5.0's confident predictions can potentially help towards a better annotation of proteomes.

Method	URL	Organism	Prediction type
SignalP 5.0	http://www.cbs.dtu.dk/services/SignalP/	A | E | P | N	S | L | T | O
SignalP 4.1[12]	http://www.cbs.dtu.dk/services/SignalP-4.1/	E | P | N	S | O
DeepSig[19]	https://deepsig.biocomp.unibo.it/deepsig/	E | P | N	S | O
LipoP[47]	http://www.cbs.dtu.dk/services/LipoP/	A | P | N	S | L | O
Philius[21]	http://www.yeastrc.org/philius/pages/philius/runPhilius.jsp	A | E | P | N	S | O
Phobius[22]	http://phobius.sbc.su.se/	A | E | P | N	S | O
PolyPhobius[48]	http://phobius.sbc.su.se/poly.html	A | E | P | N	S | O
PRED-LIPO[30]	http://bioinformatics.biol.uoa.gr/PRED-LIPO/	A | P | N	S | L | O
PRED-SIGNAL[20]	http://bioinformatics.biol.uoa.gr/PRED-SIGNAL/	A | E | P | N	S | O
PRED-TAT[25]	http://www.compgen.org/tools/PRED-TAT/	A | E | P | N	S | T | O
PrediSi[49]	http://www.predisi.de/	E | P | N	S | O
Signal-3L 2.0[33]	http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/	E | P | N	S | O
Signal-BLAST[17]	http://sigpep.services.came.sbg.ac.at/signalblast.html	A | E | P | N	S | O
Signal-CF[34]	http://www.csbio.sjtu.edu.cn/bioinf/Signal-CF/	E | P | N	S | O
SOSUIsignal[50]	http://harrier.nagahama-i-bio.ac.jp/sosui/sosuisignal/sosuisignal_submit.html	E | P | N	S | O
SPEPlip[35]	http://gpcr.biocomp.unibo.it/cgi/predictors/spep/pred_spepcgi.cgi	E | P | N	S | L | O
SPOCTOPUS[23]	http://octopus.cbr.su.se/	A | E | P | N	S | O
TATFIND[26]	http://signalfind.org/tatfind.html	A | E | P | N	T | O
TatP[51]	http://www.cbs.dtu.dk/services/TatP/	A | P | N	T | O
TOPCONS2[24]	http://www.topcons.net/	A | E | P | N	S | O

Supplementary Table 1. The full list of methods used for benchmarking of SignalP 5.0. In the 'Organism' column, where we specify the type of organisms each method was designed for or can work with, 'A' stands for Archaea, 'E' for Eukaryotes, 'N' for Gram-negative bacteria and 'P' for Gram-positive bacteria. In the 'Prediction type' column, where we specify which type of predictions each method can produce, 'S' stands for Sec/SPI signal peptides, 'L' for Lipoprotein (Sec/SPII) signal peptides, 'T' for Tat/SPI signal peptides and 'O' for other (globular and/or transmembrane proteins).

Type	Archaea	Eukaryotes	Gram-negative bacteria	Gram-positive bacteria
Sec/SPI signals	60 (50)	2,614 (210)	509 (90)	189 (25)
Sec/SPII signals	28 (19)	--	1,063 (442)	449 (201)
Tat/SPI signals	27 (22)	--	334 (98)	95 (74)
Globular	78 (63)	13,612 (6,929)	202 (103)	140 (64)
TM	44 (28)	1,044 (318)	220 (50)	50 (25)
Total	237 (182)	17,270 (7,457)	2,328 (783)	923 (389)

Supplementary Table 2. The composition of the training and test datasets used during the development of SignalP 5.0 with the benchmark dataset numbers in parentheses

SignalP 5.0 (cross-val)	Archaea		Gram-negative bacteria		Gram-positive bacteria		Eukaryotes
	MCC^1	MCC^2	MCC^1	MCC^2	MCC^1	MCC^2	MCC
Sec/SPI	0.938	0.933	0.907	0.918	0.890	0.882	0.966
Sec/SPII	0.956	0.938	0.960	0.960	0.957	0.957	-
Tat/SPI	0.977	0.958	0.981	0.981	0.868	0.866	-

Supplementary Table 3. Cross-validated performance of SignalP 5.0 on SP detection over the whole training dataset (20,758 proteins). 'MCC^1' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by TM+Globular proteins only; 'MCC^2' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by Sec/SPII SPs,Tat/SPI SPs, TM and Globular proteins.

SignalP 5.0 (cross-val)	Sec/SPI	Sec/SPII	Tat/SPI	Other
Total dataset (real/predicted)				
Sec/SPI	3,254 (339)	22 (7)	8 (1)	88 (28)
Sec/SPII	24 (14)	1,503 (637)	2 (2)	11 (9)
Tat/SPI	13 (9)	8 (6)	433 (177)	2 (2)
Globular	40 (21)	1 (1)	0 (0)	13,991 (7,137)
TM	64 (17)	1 (0)	2 (1)	1,291 (403)
Archaea (real/predicted)				
Sec/SPI	56 (46)	1 (1)	0 (0)	3 (3)
Sec/SPII	1 (1)	26 (17)	1 (1)	0 (0)
Tat/SPI	0 (0)	0 (0)	26 (21)	1 (1)
Globular	0 (0)	0 (0)	0 (0)	78 (63)
TM	1 (1)	0 (0)	0 (0)	43 (27)
Eukaryotes (real/predicted)				
Sec/SPI	2,550 (194)	-	-	64 (16)
Globular	39 (21)	-	-	13,573 (6,908)
TM	49 (13)	-	-	995 (305)
Gram-negative bacteria (real/predicted)				
Sec/SPI	474 (76)	18 (6)	3 (0)	14 (8)
Sec/SPII	19 (9)	1,040 (430)	1 (1)	3 (2)
Tat/SPI	2 (2)	3 (2)	329 (94)	0 (0)
Globular	0 (0)	1 (1)	0 (0)	201 (102)
TM	9 (2)	1 (0)	2 (1)	208 (47)
Gram-positive bacteria (real/predicted)				
Sec/SPI	174 (23)	3 (0)	5 (1)	7 (1)
Sec/SPII	4 (4)	437 (190)	0 (0)	8 (7)
Tat/SPI	11 (7)	5 (4)	78 (62)	1 (1)
Globular	1 (0)	0 (0)	0 (0)	139 (64)
TM	5 (1)	0 (0)	0 (0)	45 (24)

Supplementary Table 4 . Confusion matrices for the different type of predictions that SignalP 5.0 makes on the training and the benchmark (in brackets) dataset.

SignalP 5.0 (cross-val)	Archaea				Gram-negative bacteria				Gram-positive bacteria				Eukaryotes			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall																
Sec/SPI	0.683	0.767	0.800	0.833	0.868	0.894	0.908	0.912	0.767	0.799	0.825	0.841	0.802	0.849	0.896	0.923
Sec/SPII	0.929	0.929	0.929	0.929	0.970	0.970	0.970	0.972	0.955	0.955	0.955	0.955	-	-	-	-
Tat/SPI	0.630	0.667	0.741	0.815	0.757	0.802	0.850	0.883	0.579	0.611	0.695	0.705	-	-	-	-
CS precision																
Sec/SPI	0.707	0.793	0.828	0.862	0.877	0.903	0.917	0.921	0.744	0.744	0.800	0.815	0.795	0.841	0.887	0.914
Sec/SPII	0.963	0.963	0.963	0.963	0.970	0.970	0.970	0.972	0.964	0.964	0.964	0.964	-	-	-	-
Tat/SPI	0.630	0.667	0.741	0.815	0.755	0.800	0.848	0.881	0.663	0.699	0.795	0.807	-	-	-	-

Supplementary Table 5. Cross-validated performance of SignalP 5.0 on CS recall and precision over the whole training dataset (20,758 proteins).

	CRF and transfer learning	Only CRF	Only transfer learning	No CRF nor transfer learning
Signal peptide accuracy	0.986	0.981	0.984	0.981
Cleavage site accuracy	0.808	0.790	0.637	0.626

Supplementary Table 6. Comparison of the model performance with and without CRF and transfer learning. Since all taxonomic groups were combined for this analysis, we report just the accuracy of SignalP 5.0 (i.e. the fraction of correctly identified signal peptides and the fraction of correctly predicted cleavage sites)

Method	Archaea		Eukaryotes	Gram-negative bacteria		Gram-positive bacteria	
	MCC^1	MCC^2	MCC	MCC^1	MCC^2	MCC^1	MCC^2
SignalP 5.0	**0.922**	**0.917**	**0.883**	**0.860**	**0.830**	0.922	**0.760**
SignalP 4.1	n.d.	n.d.	0.808	0.851	0.248	**0.949**	0.148
DeepSig	n.d.	n.d.	0.819	0.792	0.166	0.870	0.115
LipoP	0.829	0.604	0.363	0.787	0.483	0.922	0.403
Philius	0.765	0.447	0.421	0.802	0.127	0.843	0.075
Phobius	0.813	0.514	0.510	0.818	0.132	0.818	0.074
PolyPhobius	0.752	0.453	0.456	0.844	0.144	0.852	0.111
PrediSi	n.d.	n.d.	0.553	0.802	0.244	0.781	0.121
PRED-LIPO	0.796	0.586	0.234	0.775	0.398	0.896	0.410
PRED-SIGNAL	0.897	0.584	0.272	0.724	0.098	0.830	0.114
PRED-TAT	0.788	0.626	0.326	0.769	0.187	0.830	0.189
Signal-3L 2.0	n.d.	n.d.	0.597	0.806	0.110	0.922	0.106
Signal-CF	n.d.	n.d.	0.326	0.561	0.106	0.558	0.084
SOSUIsignal	n.d.	n.d.	0.375	0.693	0.108	0.722	0.047
SPEPlip	n.d.	n.d.	0.655	0.746	0.498	0.646	0.350
SPOCTOPUS	0.765	0.408	0.492	**0.860**	0.127	0.922	0.109
TOPCONS2	0.765	0.432	0.477	0.860	0.131	0.897	0.071

Supplementary Table 7. Benchmarking of Sec/SPI signal peptide detection predictions. The highest performance values have been highlighted in bold. 'MCC^1' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by TM+Globular proteins only; 'MCC^2' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by Sec/SPII SPs,Tat/SPI SPs, TM and Globular proteins.

Method	Archaea				Eukaryotes				Gram-negative bacteria				Gram-positive bacteria			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall																
SignalP 5.0	0.660	0.740	0.780	0.820	**0.729**	**0.762**	**0.795**	**0.833**	**0.733**	**0.767**	0.800	0.800	**0.840**	**0.840**	**0.880**	**0.880**
SignalP 4.1	n.d.	n.d.	n.d.	n.d.	0.695	0.729	0.762	0.786	0.644	0.711	0.733	0.744	**0.840**	**0.840**	0.840	0.840
DeepSig	n.d.	n.d.	n.d.	n.d.	0.624	0.652	0.690	0.724	0.600	0.656	0.667	0.678	0.760	0.760	0.840	0.840
LipoP	0.480	0.620	0.660	0.720	0.343	0.386	0.419	0.448	0.733	**0.767**	0.789	0.789	0.600	0.600	0.640	0.640
Philius	0.580	0.680	0.700	0.700	0.619	0.686	0.743	0.781	0.700	0.744	0.789	0.811	0.600	0.600	0.600	0.600
Phobius	0.540	0.640	0.660	0.700	0.667	0.700	0.738	0.786	0.644	0.722	0.789	0.811	0.600	0.600	0.600	0.600
PolyPhobius	0.560	0.680	0.680	0.700	0.681	0.733	0.776	**0.833**	0.644	0.733	0.811	0.822	0.680	0.680	0.720	0.720
PrediSi	n.d.	n.d.	n.d.	n.d.	0.652	0.695	0.719	0.767	0.722	0.789	0.811	0.822	0.640	0.640	0.760	0.800
PRED-LIPO	0.480	0.600	0.660	0.680	0.095	0.114	0.152	0.181	0.467	0.522	0.567	0.600	0.760	0.760	0.760	0.760
PRED-SIGNAL	**0.800**	**0.900**	**0.900**	**0.900**	0.224	0.290	0.329	0.362	0.444	0.522	0.622	0.644	0.680	0.680	0.720	0.720
PRED-TAT	0.580	0.720	0.800	0.820	0.410	0.510	0.571	0.614	0.711	0.767	0.800	0.822	0.720	0.720	0.760	0.760
Signal-3L 2.0	n.d.	n.d.	n.d.	n.d.	0.648	0.686	0.733	0.762	0.644	0.700	0.722	0.733	0.800	0.800	0.840	0.840
Signal-CF	n.d.	n.d.	n.d.	n.d.	0.652	0.676	0.724	0.762	0.689	0.711	0.744	0.778	0.720	0.720	0.800	0.800
SOSUIsignal	n.d.	n.d.	n.d.	n.d.	0.176	0.329	0.467	0.576	0.267	0.367	0.567	0.622	0.200	0.240	0.280	0.440
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.710	0.733	0.771	0.810	0.611	0.678	0.722	0.733	0.680	0.680	0.720	0.720
SPOCTOPUS	0.340	0.480	0.520	0.560	0.390	0.533	0.686	0.757	0.467	0.689	**0.833**	**0.867**	0.640	0.760	0.800	0.880
TOPCONS2	0.480	0.600	0.620	0.640	0.371	0.505	0.638	0.729	0.544	0.622	0.733	0.767	0.240	0.320	0.400	0.440
CS precision																
SignalP 5.0	**0.771**	**0.688**	**0.812**	**0.812**	**0.671**	**0.702**	**0.732**	**0.732**	**0.742**	**0.775**	**0.809**	**0.809**	**0.600**	**0.600**	**0.629**	**0.629**
SignalP 4.1	n.d.	n.d.	n.d.	n.d.	0.613	0.643	0.672	0.693	0.151	0.167	0.172	0.175	0.083	0.083	0.083	0.083
DeepSig	n.d.	n.d.	n.d.	n.d.	0.604	0.631	0.668	0.700	0.131	0.144	0.146	0.148	0.073	0.073	0.080	0.080
LipoP	0.484	0.375	0.516	0.562	0.159	0.178	0.194	0.207	0.327	0.342	0.351	0.351	0.153	0.153	0.163	0.163
Philius	0.425	0.362	0.438	0.438	0.151	0.168	0.182	0.191	0.106	0.112	0.119	0.122	0.054	0.054	0.054	0.054
Phobius	0.395	0.333	0.407	0.432	0.226	0.237	0.250	0.267	0.098	0.110	0.120	0.124	0.054	0.054	0.054	0.054
PolyPhobius	0.395	0.326	0.395	0.407	0.176	0.190	0.201	0.216	0.097	0.110	0.122	0.124	0.060	0.060	0.063	0.063
PrediSi	n.d.	n.d.	n.d.	n.d.	0.273	0.291	0.301	0.321	0.144	0.157	0.162	0.164	0.062	0.062	0.074	0.078
PRED-LIPO	0.455	0.364	0.5	0.515	0.069	0.083	0.110	0.131	0.212	0.237	0.258	0.273	0.216	0.216	0.216	0.216
PRED-SIGNAL	0.489	0.435	0.489	0.489	0.066	0.085	0.096	0.106	0.076	0.089	0.106	0.110	0.060	0.060	0.064	0.064
PRED-TAT	0.493	0.397	0.548	0.562	0.080	0.099	0.111	0.119	0.125	0.135	0.141	0.145	0.082	0.082	0.087	0.087
Signal-3L 2.0	n.d.	n.d.	n.d.	n.d.	0.322	0.341	0.365	0.379	0.113	0.123	0.127	0.129	0.074	0.074	0.078	0.078
Signal-CF	n.d.	n.d.	n.d.	n.d.	0.105	0.109	0.117	0.123	0.102	0.105	0.110	0.115	0.059	0.059	0.065	0.065
SOSUIsignal	n.d.	n.d.	n.d.	n.d.	0.037	0.069	0.098	0.121	0.040	0.055	0.086	0.094	0.018	0.021	0.025	0.039
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.366	0.378	0.398	0.418	0.276	0.307	0.327	0.332	0.187	0.187	0.198	0.198

SPOCTOPUS	0.293	0.207	0.317	0.341	0.120	0.164	0.211	0.233	0.067	0.098	0.119	0.124	0.056	0.066	0.070	0.077
TOPCONS2	0.366	0.293	0.378	0.390	0.107	0.146	0.184	0.210	0.081	0.093	0.110	0.115	0.022	0.029	0.036	0.039

Supplementary Table 8. Benchmarking of Sec/SPI signal peptide cleavage site predictions, measured as recall and precision. The highest performance values have been highlighted in bold.

Method	Archaea		Gram-negative bacteria		Gram-positive bacteria	
	MCC^1	MCC^2	MCC^1	MCC^2	MCC^1	MCC^2
SignalP 5.0	**0.936**	**0.910**	**0.945**	**0.946**	**0.917**	**0.923**
LipoP	0.836	0.755	0.791	0.833	0.798	0.822
PRED-LIPO	0.766	0.743	0.629	0.707	0.775	0.775
SPEPlip	n.d.	n.d.	0.857	0.884	0.845	0.843

Supplementary Table 9. Benchmarking of Sec/SPII signal peptide detection predictions. The highest performance values have been highlighted in bold. 'MCC^1' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPII SPs and the negative dataset by TM+Globular proteins only; 'MCC^2' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPII SPs and the negative dataset by Sec/SPI SPs,Tat/SPI SPs, TM and Globular proteins.

Method	Archaea				Gram-negative bacteria				Gram-positive bacteria			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall												
SignalP 5.0	**0.895**	**0.895**	**0.895**	**0.895**	**0.964**	**0.964**	**0.964**	**0.968**	**0.925**	**0.925**	**0.925**	**0.925**
LipoP	0.684	0.684	0.737	0.737	0.860	0.860	0.860	0.862	0.831	0.831	0.831	0.831
PRED-LIPO	0.632	0.632	0.632	0.632	0.717	0.717	0.717	0.719	0.816	0.816	0.816	0.816
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.912	0.912	0.914	0.914	0.876	0.876	0.876	0.876
CS precision												
SignalP 5.0	**0.944**	**0.944**	**0.944**	**0.944**	**0.970**	**0.970**	0.970	**0.975**	**0.959**	**0.959**	**0.959**	**0.959**
LipoP	0.765	0.765	0.824	0.824	0.969	0.969	0.969	0.972	0.944	0.944	0.944	0.944
PRED-LIPO	0.923	0.923	0.923	0.923	0.969	0.969	0.969	0.972	0.921	0.921	0.921	0.921
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.969	0.969	**0.971**	0.971	0.936	0.936	0.936	0.936

Supplementary Table 10. Benchmarking of Sec/SPII signal peptide cleavage site predictions, measured as recall and precision. The highest performance values have been highlighted in bold.

Method	Archaea		Gram-negative bacteria		Gram-positive bacteria	
	MCC^1	MCC^2	MCC^1	MCC^2	MCC^1	MCC^2
SignalP 5.0	**0.972**	**0.948**	**0.958**	**0.965**	0.859	**0.889**
PRED-TAT	**0.972**	**0.948**	**0.958**	0.948	**0.903**	0.853
TatP	0.824	0.667	0.787	0.689	0.752	0.680
TATFIND	**0.972**	0.902	0.893	0.910	0.793	0.800

Supplementary Table 11. Benchmarking of Tat/SPI signal peptide detection predictions. The highest performance values have been highlighted in bold. 'MCC^1' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Tat/SPI SPs and the negative dataset by TM+Globular proteins only; 'MCC^2' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Tat/SPI SPs and the negative dataset by Sec/SPI SPs, Sec/SPII SPs, TM and Globular proteins.

Method	Archaea				Gram-negative bacteria				Gram-positive bacteria			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall												
SignalP 5.0	**0.591**	**0.636**	**0.727**	**0.773**	0.684	0.724	0.745	0.776	0.595	**0.622**	**0.676**	**0.689**
PRED-TAT	0.500	0.545	0.636	0.636	**0.735**	**0.755**	**0.776**	**0.806**	**0.622**	**0.622**	0.635	**0.689**
TatP	0.318	0.409	0.500	0.500	0.653	0.673	0.694	0.704	0.446	0.473	0.514	0.581
TATFIND	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
CS precision												
SignalP 5.0	**0.591**	**0.636**	**0.727**	**0.727**	0.698	**0.740**	**0.760**	0.760	**0.698**	**0.730**	**0.794**	**0.794**
PRED-TAT	0.500	0.545	0.636	0.636	**0.713**	0.733	0.752	**0.782**	0.590	0.590	0.603	0.654
TatP	0.269	0.346	0.423	0.423	0.427	0.440	0.453	0.460	0.355	0.376	0.409	0.462
TATFIND	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.

Supplementary Table 12. Benchmarking of Tat/SPI signal peptide cleavage site predictions, measured as recall and precision. The highest performance values have been highlighted in bold.